ARMY RESEARCH LABORATORY

Evidence for Increased Discriminability in Judging the Acceptability of Machine Translations: The Case for Magnitude Estimation

by James D. Walrath

ARL-MR-0720 **May 2009**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-MR-0720 **May 2009**

Evidence for Increased Discriminability in Judging the Acceptability of Machine Translations: The Case for Magnitude Estimation

James D. Walrath
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
May 2009	Progress	1ST Q FY09

4. TITLE AND SUBTITLE
Evidence for Increased Discriminability in Judging the Acceptability of Machine Translations: The Case for Magnitude Estimation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
James D. Walrath

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: AMSRD-ARL-CI-IT
2800 Powder Mill Road
Adelphi MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-MR-0720

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
An earlier experiment, in which magnitude estimation (ME) was used as the method for judging the acceptability of machine translations, was replicated with one exception. The current study used a four-point Likert scale as the measurement methodology. In the earlier work, using ME, judges easily discriminated between two sets of machine translations, one containing 25% more correctly translated names than the other. In this study, using a Likert scale, judges were not able to make the same discrimination. These results support the theory that ME may be a superior measurement methodology for assessing the acceptability of machine translations.

15. SUBJECT TERMS
Acceptability of machine translation, magnitude estimation, Likert scale

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			James D. Walrath
Unclassified	Unclassified	Unclassified	UU	20	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-5616

Contents

INTENTIONALLY LEFT BLANK.

1. Introduction

Our commercial, political, and military global interactions have broadened to include countries that are home to low resource languages such as Dari, Pashto, and Swahili. One result of this is a deepening interest in the field of machine translation (MT). There is simply too much written and spoken in other languages, particularly in these low resource languages, than can be dealt with by human translators. Responding to the need for translation services with many languages and in many formats, scientists in industry, academia, and government are generating a myriad of MT solutions, each with its own incremental improvements.

As the MT field has grown, so has the challenge in evaluating and selecting systems. Computational linguists, having principal stewardship over MT development, have also set out to create computational methods of measuring the "goodness" of these systems. Several of these metrics have received at least tacit acceptance in the community. For many, however, the gold standard remains the judgment of certified bilingual interpreters, such as those at the Defense Language Institute in Monterey, CA. These professionals are often asked to judge the acceptability of machine translations, one sentence or utterance at a time. Their many judgments become the data that are used to draw inferences about the desirability of one MT system compared to another, or the degree of success a new version enjoys over an older version. These are very important measurements because they drive purchase decisions, which put systems in Soldiers' hands, and the Soldiers should have the best systems available.

When considering the relation between the physical world and our perception of it, one fundamental perceptual question is how do we judge the magnitude of a given stimulus parameter (e.g., translation acceptability) and thus how do we judge the degree of similarity or difference between stimuli? In the MT field, the traditional methodology used to record these judgments has often been some variation of an ordinal scale, usually a Likert scale, after Rensis Likert who first published research making use of the methodology (Likert, 1932). The Likert scale consists of a number of statements describing some attribute (e.g., the acceptability of a machine translation) and requires the participant to pick the statement that best describes their judgment of that attribute (see appendix B for an example).

However, there are shortcomings when using the Likert scale in some applications. First, many feel that data from Likert scales, being ordinal, is not amenable to parametric statistical analysis, because the mean and standard deviation cannot be used as measures of central tendency and variability, respectively. This limits the statistical techniques that can be brought into play when analyzing the data. A second issue, and the one central to this report, is that Likert scales may not allow judges the full range of discriminability of which they are capable. We have all experienced Likert scales (think, surveys) that have frustrated us due to a lack of choices.

Evaluations have often used four-point Likert scales when asking translation professionals to rate the adequacy of a machine translation. As will become evident in this report, people are capable of expressing many more than four levels of discrimination when making this judgment. We can capitalize on this ability by using a psychophysical methodology that accommodates these many levels of expression. Once such methodology is magnitude estimation (ME).

ME is a method of psychophysical ratio scaling developed by S. S. Stevens (1956) in the mid-20th century and has been frequently used in investigations as diverse as judging the brightness of a light or the pitch of a tone to the prestige of occupations or the goodness of moral judgments. ME requires observers to make direct numerical estimates of the magnitude of a stimulus characteristic. Observers may use any positive, non-zero number. Often these estimates of stimulus magnitude are made with reference to a standard stimulus, presented first, with a magnitude already assigned to it by the experimenter. The observer is asked to consider the score given to the standard stimulus (or modulus) when deciding on a score for each element of the test set.

ME has been successfully used in evaluating linguistic acceptability (Bard et al., 1996). It has also been employed to measure translation adequacy. Investigating the differential effect of correct name translation on human and automated judgments of translation adequacy, Vanni and Walrath (2008) used ME to score judge's acceptability of machine translated sentences. Arabic sentences were translated into English, forming the Control Stimulus Set. An Enhanced Stimulus Set was then created by increasing the number of correct name translations by 25%. Since judges were monolingual English speakers, a reference translation of each of the Arabic sentences was provided. One-half of the judges compared the English machine translations from the Control Set to their respective reference translations, while the remaining judges compared the translations in the Enhanced Set to the reference translations. Judges were asked to score the degree to which the machine translation conveyed the meaning present in the reference translation. Automated metrics were also collected for both sets of translations. As one might expect, the Enhanced Set was judged to be significantly more acceptable than the Control Set, both by the human judges and by the automated metrics. Of importance to Vanni and Walrath was the fact that the benefit offered by improved name translation was far greater for the human than for the automated metrics, indicating that correct name translation has a cognitive gravitas not correctly modeled by the automated methods.

For this report, the importance of Vanni and Walrath's work is that judges, using ME, had no trouble differentiating between the Control and Enhanced Sets of translations. Specifically, the difference in scores between the two groups was statistically significant.

The research reported here looked to see if a Likert scale methodology would also result in a significant difference in the judges' scoring of the two Stimulus Sets.

2. Method

The methods used in this research were identical with those used in the Vanni and Walrath (2008) work, except here a Likert scale was used rather than ME. Readers are directed to their report for full details.

Briefly, 20 Arabic sentences were translated into English using a research grade text-to-text MT system. These sentences were selected from open source material assembled in support of an annual MT competition.[1] These translations became the Control Stimulus Set, and contained 76 instances of incorrect name translation. Nineteen (25%) of these were randomly selected and correctly translated. This set formed the Enhanced Stimulus Set.

Because the judges were monolingual English speakers, a reference translation of each of the 20 Arabic sentences was also created by a professional human translator. One-half of the judges compared the segments from the Control Set with the reference translations, while the remaining subjects compared the segments from the Enhanced Set with the reference translations. Subjects were asked to judge the degree to which the machine translation conveyed the meaning present in the reference translation.

2.1 Judges

Nine adult males and one adult female volunteered for participation in this study; they were non-linguists and all were employed by the U.S. Department of Defense. No compensation was received for participation in the study, nor did any of the judges have prior experience evaluating the acceptability of machine translations. Judges were randomly assigned to one of two groups. The control group was presented with the Control Set of machine translations, while the experimental group saw the Enhanced Set of machine translations.

2.2 Apparatus

Written instructions (appendix A) and test booklets (example in appendix B) were prepared. The instructions contained example translations that were fabricated by the experimenter to assist in training. Each test booklet contained 20 written machine translated sentences appropriate to its group assignment. As described previously and illustrated in the example in appendix B, each translation was accompanied by its reference translation. Judges used a pen or pencil to record their scores.

[1] The National Institute of Standards and Technology (NIST) annually conducts a competition among MT research systems. These data were part of the NIST 2008 Open MT Evaluation. For more information on this program, see http://www.nist.gov/speech/tests/mt/2008/doc.

2.3 Procedure

Judges participated at their convenience in their offices. After reading the instructions, the experimenter answered any remaining questions. Judges then received their test booklet and were left alone to complete the task. Upon completion, judges returned the instructions and test booklet to the experimenter.

Judges were asked to consider how much of the meaning present in the reference translation was also present in the machine translation. Thus, translation acceptability was defined in terms of meaning. Judges expressed their degree of acceptability, for each translation, by circling the number on the Likert scale most closely matching their judgment of the translation's acceptability.

3. Results

Recall that the study by Vanni and Walrath (2008), using ME, found the Enhanced Stimulus Set of machine translations to be significantly more acceptable than the Control Stimulus Set, t= –2.685 with 38 degrees of freedom (P=.011).

Here, however, the Likert measurement method failed to find a significant difference between the same two groups. The median scores for the 20 Control Set segments and the 20 Enhanced Set segments were calculated. For both sets, the overall median was 2. A Mann-Whitney Rank Sum Test failed to find a statistically significant difference between the groups, T= 463.5 (P=.119).

Likert scale data are often collapsed to nominal form by categorizing responses as either "acceptable" or "unacceptable" (e.g., a score of 3 or 4 is scored as acceptable, 1 or 2 is scored as unacceptable). A Chi-Square test is then applied to the transformed data. This analysis was performed on these data and, again, no significant difference between groups was found, Chi-square = 1.021 with 1 degree of freedom (P=0.312).

4. Discussion

The objective of this research was to determine if a four-point Likert scale could offer the same level of discriminability as ME when judging the acceptability of machine translations. The Likert scale methodology was clearly inferior to ME for the sets of translations used in these experiments. The results lend support to the argument that ME methodology allows for greater discriminability with which to measure translation acceptability. This heightened discriminability of ME seems reasonable when considering the levels of expression used by the judges in both studies. Obviously the judges using the Likert scale were constrained to four

levels of expression (i.e., they could choose one of four positions on the scale). The judges using ME scored their judgments by writing down any non-zero positive number. They used, on average, 10.4 different numbers in scoring the 20 translations (i.e., they averaged 10.4 levels of expression), two and one-half times more than allowed by the Likert scale. Thus ME offered the opportunity for making finer grained judgments of the acceptability of machine translations, and the judges took advantage of that opportunity, even though many of them felt they wouldn't be able to do ME. Bard et al. (1996) said the following about their experience using ME in a linguistic acceptability study: "Whatever subjects do when magnitude-estimating linguistic acceptability, and however odd they find the whole process at first, they clearly have this ability in their psychological repertoire, just as they have the ability to give proportionate judgments of brightness or prestige." (p. 60)

5. Conclusions

It is tempting to generalize this finding to any set of translations from any MT engine in any language, but these data cannot support such generalizations. Even so, the fact that judges working with ME used so many more levels of expression than would be tenable with a Likert scale is compelling evidence supporting the theory of ME's general superiority.

Further research, using different languages and translation systems would be helpful in accepting or rejecting the theory of ME's general superiority in this kind of work. For example, do the same apparent advantages ME enjoys over Likert scales hold true for speech-to-speech MT systems? Further, neither the current work nor the Vanni and Walrath (2008) study used bilingual judges who can directly compare the MT input language text to the output text. Thus, there are experiments yet to be done, but the future for ME appears bright.

6. References

Bard, E. G.; Robertson, D.; Sorace, A. Magnitude Estimation of Linguistic Acceptability. *Language* **1996**, *72* (1), 32–68.

Grescheider, G. *Psychophysics Method, Theory, and Application*, 2nd ed.; Hillsdale: New Jersey: Lawrence Erlbaum Associates, 1985.

Likert, R A Technique for the Measurement of Attitudes. *Archives of Psychology* **1932**, *140*, 1–55.

Stevens, S. S. The Direct Estimation of Sensory Magnitude—Loudness. *American Journal of Psychology* **1956**, 69, 1–25.

Vanni, M.; Walrath, J. *Differential Effect of Correct Name Translation on Human and Automated Judgments of Translation Acceptability: A Pilot Study*; ARL-TR-4630; U. S. Army Research Laboratory: Adelphi, MD, 2008.

Appendix A. Instructions to the Judges

Appendix A includes the instructions to the judges.

Thank you for taking the time to help improve machine translation.

You will be asked to read sentences that have been translated from Arabic to English using a machine. Each machine translation will be accompanied by a translation of the same Arabic sentence but done by a certified bilingual human translator and is considered the translation "gold standard." So, each sentence in Arabic is translated by the human translator and by a machine translation system. You will see both of these translations. Your task is to judge how the machine translation compares to the human translation.

Of interest is the degree to which the machine translation conveys the meaning present in the human translation. The machine translation may not contain good, natural-sounding English like the human translation but you need to overlook that. The question to ask yourself is, "Do I get the same meaning from the machine translation as I do from the human translation?"

Let's look at some examples.

Human Translation:

Mr. Goldman visited his uncle Ralph on Tuesday in Paris.

Machine translation:

Tuesday, Mr. Gold in Paris to visit his uncle, Ralph.

In this example, most all of the meaning available in the human translation is also available in the machine translation. "Mr. Goldman" is incorrectly translated as "Mr. Gold." The human translation is in the past tense and the machine translation is in either the present or future tense. On balance, though, nearly all the meaning survives the machine translation. The readability of the machine translation is not great but, again, we want you to ignore that.

In brief, the pros and cons of this translation are:

Pros: "uncle Ralph," "Tuesday," and "Paris" are all correctly translated

Cons: "Mr. Goldman" is incorrectly translated as "Mr. Gold"

Let's look at another example.

Human translation:

When the 82nd Airborne jumped at Market Garden, General Gavin was the first one out of the plane.

Machine translation:

82 surge in the market when the Hanging Gardens, General Gavin is the first one out of the plane.

Here less information survives the machine translation. The fact that General Gavin jumped out of the plane first is in the machine translation even though the tense has been changed from past to present. However, "82nd Airborne" and "Market Garden" have been lost. A student of World War II history may be able to make sense of the machine translation but the reader should be able to understand the meaning of the translation without any special knowledge.

Pros: "General Gavin" is correctly translated; what General Gavin did is correctly translated

Cons: "82nd Airborne" and "Market Garden" are not correctly translated

Another example.

Human translation:

Major Hassan reported to Colonel Ali that a dozen Humvees located in Al Asad Base aren't ready.

Machine translation:

Transfer to Colonel Hassan leading to a dozen cars Alhmralamugodh base Assad not ready.

This machine translation gets many things wrong. The person "Hassan" survives the translation but "Colonel Ali" does not. The rank of Hassan is changed from Major to Colonel. It seems that 12 cars (that are actually Humvees) are being transferred to (now) Colonel Hassan—a meaning not in the human translation. We have no idea what Alhmralamugodh is. There is a reference to base Assad (a mistranslation of Al Asad) not being ready when, in truth, the vehicles aren't ready, not the base.

Pros: The name "Hassan" survives translation; 12 vehicles, of some description, are mentioned

Cons: Hassan's rank should be Major, not Colonel; "Colonel Ali" and "Humvees" are not translated; "Al Asad" is translated as "Assad" (similar but different); the machine translation refers to a "transfer" which is not mentioned in the human translation.

As you can see from these three actual examples, the amount of meaning retained in a machine translation can vary widely. So how are you to assign a value to each sample machine translation? The answer follows.

To score the machine translation, ask yourself this question: How much of the meaning present in the human translation is also present in the machine translation? Is it adequate or not?

If you judge the translation to have **adequate meaning**, score it either **4** (completely adequate) or **3** (mostly adequate).

If you judge the translation to have **inadequate meaning**, score it either **2** (mostly inadequate) or **1** (completely inadequate).

To further explain consider the following examples. The first two would be judged as having adequate meaning and the last to as having inadequate meaning.

Score = 4 The meaning in the translation is **completely adequate**.

Example: Human Translation: Cars were checked for weapons.

Machine Translation: Cars had checks of weapons.

Score = 3 The translation is **mostly adequate**.

Example: Human Translation: We were told to go outside the house.

Machine Translation: We commanded have leaving outside.

Score = 2 The translation is **mostly inadequate.**

Example: Human Translation: My father, my mother, and my brothers were here.

Machine Translation: At the end who and brother.

Score = 1 The translation is **completely inadequate**.

Example: *Human Translation:* Coalition forces found weapons in his car.
Machine *Translation*: Uncle here melon on grandfather to go.

Following are 20 machine translations with their associated human translations. Circle the number that best describes how much of the meaning in the human translation is contained in the machine translation.

INTENTIONALLY LEFT BLANK.

Appendix B. A Sample from a Judge's Test Booklet

Appendix B contains one page from the judge's test booklet.

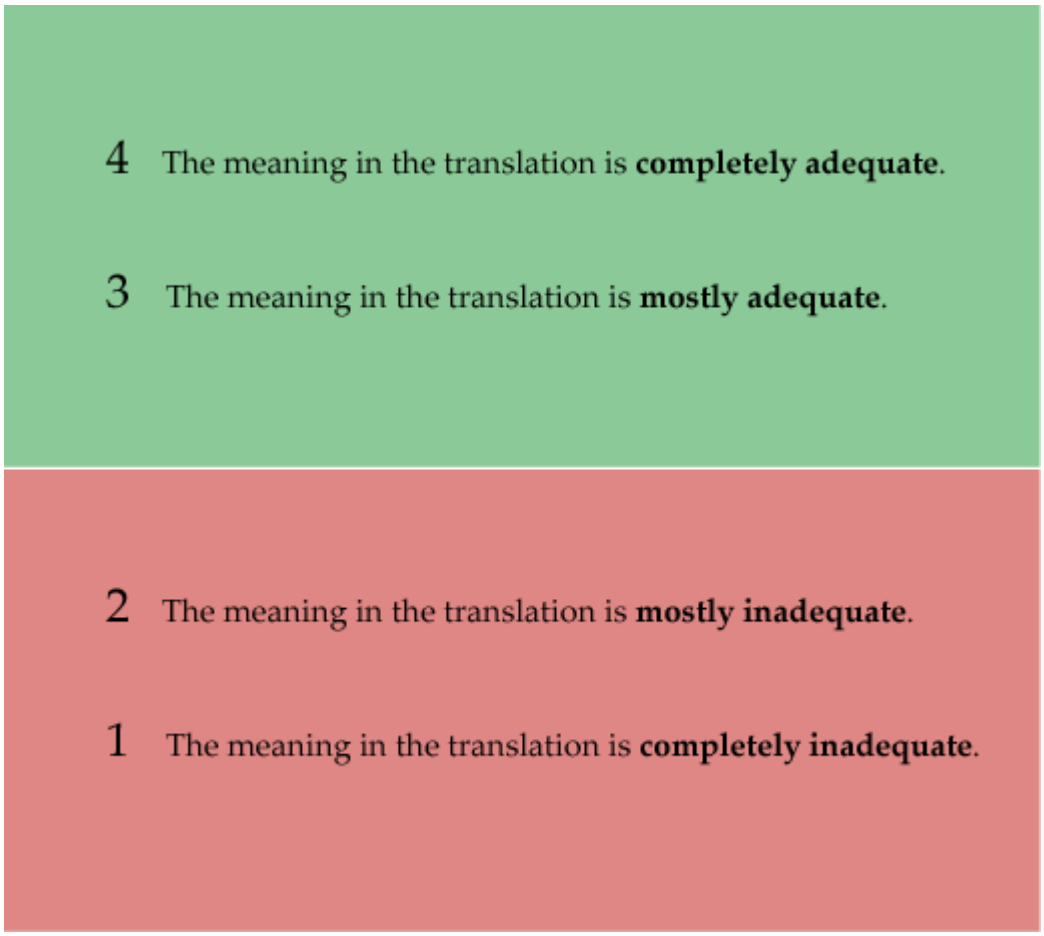
Human translation:

Martin Jager, spokesperson for the German Foreign Ministry, said that two Germans are missing in Afghanistan.

Machine translation:

Confirmed passer-by figs ga pulls, the spokesman the foreign ministry A only that in the citizens the german lost two in Afghanistan.

Circle the number that represents how much of the meaning present in the human translation is also present in the machine translation.



INTENTIONALLY LEFT BLANK.

NO. OF COPIES	ORGANIZATION
1 PDF	ADMNSTR DEFNS TECHL INFO CTR 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1 CD	OFC OF THE SECY OF DEFNS ATTN ODDRE (R&AT) THE PENTAGON WASHINGTON DC 20301-3080
1 HC	US GOVERNMENT PRINT OFF DEPOSITORY RECEIVING SECTION ATTN MAIL STOP IDAD J TATE 732 NORTH CAPITAL ST NW WASHINGTON DC 20402
1 HC	US ARMY RSRCH LAB ATTN AMSRD ARL CI OK TP TECHL LIB T LANDFRIED BLDG 4600 ABERDEEN PROVING GROUND MD 21005-5066
17 HCs 1 PDF	US ARMY RSRCH LAB ATTN AMSRD ARL CI I B BROOME (1 HC) ATTN AMSRD ARL CI IT JD WALRATH (12 HCs, 1PDF) ATTN AMSRD ARL CI IT V M HOLLAND (1 HC) ATTN AMSRD ARL CI OK PE TECHL PUB (1 HC) ATTN AMSRD ARL CI OK TL TECHNICAL LIBRARY (1 HC) ATTN IMNE ALC IMS MAIL & RECORDS MGMT (1 HC) ADELPHI MD 2078

TOTAL: 22 (1 CD, 19 HCs, 2 PDF)